Backward Induction with Players who Doubt Others' Faultlessness*

Aviad Heifetz† and Ady Pauzner‡

This version: May 2005

Abstract

We investigate the robustness of the backward-induction outcome, in binary-action extensive-form games, to the introduction of small mistakes in reasoning. Specifically, 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 show that, in a long centipede game, the prediction that players do not cooperate fails under this perturbation. Importantly, this result does not depend on forward induction or reputation reasoning. It particular, it applies to finite horizon overlapping generations models with fiat money.

J.E.L Classification No. C73, Field: Game Theory

---

*We thank the associate editor and two anonymous referees for very helpful comments. We also benefited from discussions with Eddie Dekel, Herakles Polemarchakis, and Asher Wolinsky.

†The Economics and Management Department, The Open University of Israel, aviadhe@openu.ac.il

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

1 Introduction

Backward induction is the predominant solution concept used to predict behavior in extensive form games of complete information. Nonetheless, there are numerous examples in which the backward induction outcome seems to contradict our "common sense" intuition as to how human agents would act.

The centipede game (Rosenthal 1981) is a particularly striking example. In this game, continual cooperation between two parties can yield substantial gains to both, while the backward induction outcome precludes cooperation altogether. Experimental evidence (see, for example, McKelvey and Palfrey (1992)) suggests that human subjects do not adhere to backward induction reasoning in this case: a significant proportion of subjects continue for several rounds. A closely related game is the finite overlapping generations model 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. Yet, it is hard to believe that dollars would not be used in trade today.

This paper claims that the backward induction solution concept sometimes hinges on the assumption that players are *absolutely* certain of their conclusions regarding others' reasoning. We consider a slight deviation from this assumption and show that, in some cases, the predictions of the model change considerably.

Specifically, we develop a model in which a player who contemplates the best action at some future decision node attributes some small but positive probability to the possibility that other players may reach a different conclusion when they carry out the same analysis. Moreover, each player believes that the other players have the same doubts, that other players believe that others maintain such doubts, and so on. In other words, common knowledge of rationality is replaced by *common certainty that with some small probability, the other players might conclude differently than oneself when they contemplate the best action at each decision node.*

Our model applies to binary-action games in agent form. Each agent is characterized by a set of "types". A type represents the way the agent thinks when she analyzes the decision problem of agents in all the nodes in the subgame that starts with her. For each

of these decision problems, the type specifies whether she is "correct" – and believes that the action with the higher payoff is the best one, or "confused" – and believes that the action with the lower payoff is the best one. In other words, a confused agent knows very well how to add and multiply numbers, but whenever she has to compare two payoffs x and y where $x > y$, she concludes by mistake that y is preferred to x.

A type is represented by a collection of binary digits (1 representing "correct" and 0 representing "confused"), one digit for each node in the player's subgame. A given type, however, does not understand the meaning of the 1's and 0's in her name. She believes that her way of analyzing the game is the objectively correct way, and believes that other agents tend to analyze the game in a way similar to hers. This is modeled by the way she assigns a probability distribution over the types of another agent: the more similar the type, the higher the probability. More precisely, one agent (of a given type) believes that the probability that the type of another agent differs from hers in a given digit is ε ($<\frac{1}{2}$), independently across digits. Thus, a confused type attributes a probability of $1-\varepsilon$ to the event that any other player who considers the same node would think like her, and a probability of ε to the event that the other player is "confused" (in her view) and compares payoffs in the objectively correct way.

When applied to the centipede game, the prediction of this model differs considerably from that of the "fully rational" model. Cooperation among the players ensues for a long while and only a few steps before the end of the game does it break down. To see why small doubts can induce cooperation, consider the following deliberation 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 if I assumed that it is commonly known that everyone 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 least at 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.

It is enough that this happens once in order for player 2 to continue: When analyzing the game backwards from the node where she reasoned differently, player 2 will then think that everyone should cooperate. (Since every mistake leads to a reversal in her perceived best action, any odd number of mistakes will also induce make player 2 to cooperate.) Since there are so many decision nodes where this might happen, albeit 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 the case that 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."[1]

A number of papers have studied the sensitivity of the backward induction outcome to slight deviations from the common knowledge of rationality assumption. Kreps, Milgrom, Roberts, and Wilson (1982) and Kreps (1990, p. 536) showed that cooperation in the centipede game or in the finitely repeated prisoners' dilemma can emerge if there is a small chance that one of the players is an irrational, cooperator type. The driving force behind these results is "forward induction" or "reputation" reasoning, in which the rational type of a player mimics her cooperator type in order to convince her opponent that she is likely to cooperate in the future. 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 assignment of beliefs after a zero probability event occurs, and thus can accommodate forward induction reasoning.

Our model, by contrast, does not rely on forward induction reasoning. The prediction that players will cooperate in the centipede game relies purely on backward induction.

[1]This type of reasoning will continue to be valid as long as the players are not too close to 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.

That is, the action that an agent chooses depends only on the analysis of the subgame that starts with her, and is in no way affected by the other parts of the game. Therefore, she can never infer, from the actions of another player at an early stage of the game, what (another agent of) that player is likely to do in the subgame. Nevertheless, our model predicts cooperation in long agent-form centipede games. Thus, our results also apply to finite overlapping generations models of fiat money. In such models, an agent accepts money only if she believes that the next agent will accept it; she does not expect her action to affect the next agent's beliefs regarding the likelihood that money will continue to be accepted in the future. While the reputation models do not apply in this case, our model shows that the conclusion that money cannot have value when the horizon is finite depends crucially on the assumption that players are sure that they know exactly how others think.

Another possible deviation from full rationality was studied by Selten (1975). Here, at each node there is a slight chance that a player might "tremble" and mistakenly play a suboptimal action. It is important to differentiate between our *mistakes in reasoning* and Selten's *mistakes in acting*. 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 ahead, there may be a relatively high probability of a mismatch of conclusions between two consecutive players regarding the best action. 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 "tremble" and 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 regardless of the length of the game.[2]

[2]Aumann (1992) constructs an example in which there is an irrational type that mistakenly continues at one of the decision nodes. The ex-ante probability of the mistaken type is very small. However, conditional on reaching that node, the probability is large enough (relative to the payoff differentials) that the best action for the player at the preceding node is to continue. This induces all players from that node backwards to continue. Note that Selten's counterpart to Aumann's assumption is that the tremble (for the player at that node) is large. The purpose of our paper is to show that small

The remainder of this paper is organized as follows. In section 2 we present the formal model for binary-action games. In section 3 we apply the model to the overlapping-generation version of the centipede game; a self-contained example is presented in Subsection 3.1. Section 4 concludes with discussions on the interpretation of our type space and of other games – the prisoners' dilemma and the two-player centipede game. Proofs are relegated to the appendix.

2 The Model

Consider a binary-action, finite, extensive-form game G with perfect information and set of players I. The game is described by a tree (Z, N, A), where Z is the set of leaves (terminal nodes), N is the set of (nonterminal) nodes, and A is the set of arcs. Elements of A are ordered pairs (n, m), and $n_* = \{m : (n, m) \in A\}$ is the (binary) set of immediate successors of n. Player i's payoff is a function $f_i : Z \to R$.

We assume that the game is played in "agent form". Thus, without loss of generality, we can identify the set of agents with the set of (non-terminal) nodes N. Agent n's set of (pure) strategies is simply the set of immediate successors n_*.

Let S_n be the set of nodes in the subgame starting at node n. If the agents in this subgame choose actions $(a_{n'})_{n' \in S_n}$, agent n obtains the payoff $u_n\left((a_{n'})_{n' \in S_n}\right)$, computed in the usual way.

$T_n = \{0,1\}^{S_n}$ is the set of types of agent n, with a typical element denoted $t_n = (t_n^{n'})_{n' \in S_n}$. The belief $b_{t_n}^{n'}$ of type t_n about the agent $n' \in S_n \setminus n$ who plays after her, is a probability measure over her set of types $T_{n'}$. Denoting $d(t_n, t_{n'}) = \#\left\{n'' \in S_{n'} : t_{n'}^{n''} \neq t_n^{n''}\right\}$ and $e(t_n, t_{n'}) = \#\left\{n'' \in S_{n'} : t_{n'}^{n''} = t_n^{n''}\right\}$, $b_{t_n}^{n'}$ is defined by:

$$b_{t_n}^{n'}\left(t_{n'}\right) = \varepsilon^{d(t_n, t_{n'})}\left(1-\varepsilon\right)^{e(t_n, t_{n'})}$$

where $\varepsilon \in \left(0, \frac{1}{2}\right)$ is the (common) probability of confusion. Thus the probability that

"trembles" (in reasoning) can accumulate in long games until the probability of continuation reaches that assumed by Aumann. (Note that Aumann's information structure is not of "agent-form", i.e., a player's type determines her action in more than one node. This, however, is not the driving force behind his cooperative result.)

t_n assigns to $t_{n'}$ increases with the number of nodes over which their types "agree." We assume that the beliefs of t_n over the types $t_{n'}$ of different agents $n' \in S_n$ are independent. The interpretation of this is that types interpret their names as *neutral tags*, and do not conceive the 0-s in their names as denoting nodes where they are "mistaken". Type t_n takes its own name as the point of reference, as if it were always right, and consider types $t_{n'}$ that are very different from itself as peculiar and rare.

Computation of the backward induction outcome in our model is done recursively. Let $U_n(t_n, a)$ denote the expected payoff of agent n, with beliefs determined by her type t_n, if she takes action $a \in n_*$, and let $a_n(t_n)$ be her preferred action. Then:

$$U_n(t_n, a) = \begin{cases} f_n(a) & \text{if} \quad a \text{ is a leaf} \\ \sum\limits_{(t_{n'})_{n' \in S_a} \in \underset{n' \in S_a}{\times} T_{n'}} \left(\prod\limits_{n' \in S_a} b_{t_n}^{n'} \left((t_{n'}) \right) \right) \cdot u_n \left(a, \left(a_{n'} \left(t_{n'} \right) \right)_{n' \in S_a} \right) & \text{if} \quad a \text{ is a node} \end{cases} \tag{I}$$

where

$$a_n(t_n) = \begin{cases} \arg\max_{a \in n_*} U_n(t_n, a) & \text{if} \quad t_n^n = 1 \\ \arg\min_{a \in n_*} U_n(t_n, a) & \text{if} \quad t_n^n = 0 \end{cases} \tag{II}$$

We restrict attention to games and ε for which the arg max and arg min in this definition are always singletons (i.e., $U_n(t_n, a)$ is not constant in $a \in n_*$). Clearly, this holds for generic games.

The recursive definition is applied as follows. For agents n at nodes whose successors are only leaves, the payoff $U_n(t_n, a)$ from playing action a is simply $f_n(a)$ (equation I). The agent has to choose between the two payoffs $\{f_n(a) : a \in n_*\}$. Type 1 chooses the "right" action, i.e., the one with the higher payoff. Type 0 chooses the wrong action, i.e., the one with the lower payoff (Equation II). Next, we move to nodes whose successors are either leaves or nodes whose successors are only leaves. For the former, payoffs are computed as before, according to the first line in Equation I. For the latter, we have already computed the chosen actions (as a function of types), and thus can apply the second line in Equation I. Now again we can employ Equation II to compute the actions. Continuing in this way, we eventually cover the whole tree.[3]

[3]If we were to extend the definition of the model to extensive-form games with more than two actions

3 Application: cooperation in the centipede game

In this section we study an application of our model to a class of centipede games. We will see that our solution concept yields different predictions when compared to the standard backward-induction solution concept. We work with an overlapping-generations version, in which the payoff structure is simple, and allows for an analytic solution.

Consider a centipede game with N players, each assigned to one of N consecutive decision nodes. We enumerate the nodes *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 can either continue or quit at her turn. The player can 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 the next player quits, she receives 0; if the next player continues, her payoff is $d > 2$. The last player gets 1 by quitting and 0 by continuing. The probability of confusion, ε, is assumed to be positive and less than $\frac{1}{d}$.

The overlapping-generations centipede game 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 utility (d) if the following agent accepted the money in exchange for her own endowment. But if the following agent declines to accept money, the utility is 0 (as paper money has no intrinsic value). It has been shown that in an infinite-horizon world, there is an equilibrium in which fiat money has value and can be used for trade (see Samuelson 1958). However, 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

---

per node, the digit of the type for each node would assume as many values as there are actions at that node. It would then be natural to assume that a type who is not confused at that node assigns an overall small probability to the (more than one) possible kinds of confusions at that node. It is less straightforward, however, to decide what probabilities each *confused* type would assign – both to the truly non confused type, and to the other kinds of confusion at that node.

this money will be useless. Continuing this backward induction reasoning, one can show that no agent would ever accept fiat money. 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.

How is our model applied to this game? Consider player n. Her set of types $T_n = \{0,1\}^{S_n}$ is the set of all n-digits binary numbers, where 0 in the k-th digit corresponds to confusion at node k. (Recall that "confusion" means a reversal of the usual ordering on numbers, i.e., concluding that payoff y is preferred to payoff x when $x > y$.).

Consider now player $m > n$. How does she reason about player n? This depends on her own type t_m. The belief of type t_m is a probability distribution over T_n. The probability it assigns to type $t_n \in T_n$ is computed by comparing the n-digit number t_n to the first n digits in t_m; this probability is $\varepsilon^\ell(1-\varepsilon)^{n-\ell}$, where ℓ is the number of digits in which the two numbers differ.

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. In our multi-player centipede game, the best action for player $m = n+1$ depends on the probability that player n continues. In the case that $n+1$ is not confused at her decision node, she will choose to continue if she believes that the probability that n will also continue 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; at that node, player 1 simply compares the payoff of quitting, 1, to that of continuing, 0). In the case that $n+1$ is confused, she will of course choose the opposite action to the one implied by the above rule.

3.1 An Example

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. Clearly, the usual backward

induction argument implies that all the players choose to exit if their node is ever reached.

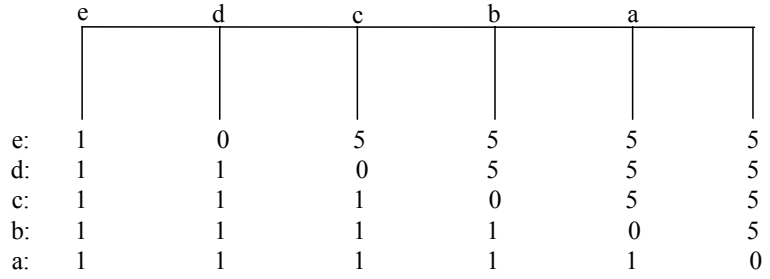

Figure 1: A 5-player centipede game

Suppose now that the probability of confusion is $\varepsilon = 0.1$. How do the players play the game? Let us analyze the game backwards. What does player b (who is not confused) think that player a will do? She believes that with probability $\varepsilon = 0.1$ player a will get confused, will assess 0 as better than 1, and will continue. If player a is not confused (probability $1 - \varepsilon = 0.9$), she will exit. 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 payoff of 1 that she gets by quitting and therefore she quits.

What does a (non-confused) 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 in computing her own best response 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 $0.82 \times 0 + 2 \times 0.1 \times 0.9 \times 5 = 0.9$. This is still less than the pay off of 1 that she can secure by quitting and therefore she quits.

What does the non-confused type of player d think that c will do? She believes that c quits unless c was 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 in all three cases (probability $0.1^3 = 0.001$). In the complementary event that c either never got confused or got confused two mutually-compensating times, c continues. Thus, if d continues, her expected payoff is

$0.757 \times 0 + 0.243 \times 5 = 1.22$, which is better than the payoff of 1 she gets by quitting. Therefore, she continues!

Finally, the non-confused type of player e is almost certain that d will continue. Why? Consider a type of d who is not confused in her own shoes. As explained above, if d never got confused in the shoes of a, b and c, she would continue. This holds also if she got confused twice when she reasoned about a, b and c. But what if d got confused once or thrice in the shoes of a, b and c? Also in this case d would continue – "by mistake": as may easily be calculated, such a type of d would ascribe probability $1 - 0.243 = 0.757$ to the event that c continues. Thus, any type of d who is not confused in her own shoes must continue; and clearly, every type of d who *is* confused in her own shoes quits. Thus, (the non-confused type of) e assigns a probability of $1 - \varepsilon = 0.9$ to the event that d continues. As a result, e will continue.

Figure 2 describes the probabilities of continuation of each player, as viewed by the non-confused type of the preceding player.

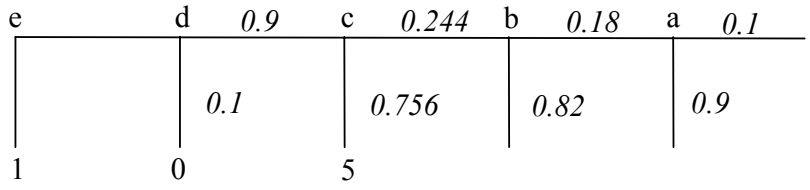

Figure 2: Probabilities of continuation.

3.2 The general result

The following theorem states that a long centipede game has a concluding segment, before which players who are not confused (i.e., of type $111\ldots1$) continue, and furthermore assign a high probability to continuation everywhere before that concluding segment.

Theorem 3.1. *There is an integer $n = n(d, \varepsilon)$, such that the types $111\ldots1$ of all players from $n+1$ on (backwards towards the beginning of the game) continue, and the types $111\ldots1$ of the players $1, \ldots n$ quit. Moreover, types $111\ldots1$ of all players $i > n$ believe that player $i-1$ continues with probability at least $1 - \frac{1}{d} - \varepsilon\left(1 - \frac{2}{d}\right)$. Finally, $n(d, \varepsilon)$ decreases in both d and ε.*

To see the intuition for the theorem, we now analyze the example in Section 3.1 and extend it by adding players f, g,... .

For each player (except player a) and each of her types, we can divide the sequence of digits that composes the type to two: one digit describes whether she is confused in her own shoes, and the other digits describes her belief regarding the types of the next player. The belief leads to an assessment of the probability that the next player quits or continues, which determines what the best action is. The digit of the player itself determines whether she indeed takes the best action, or mistakenly takes the wrong action. As for the last player a, since there are no further players, the best action for her is simply determined by the two terminal payoffs – and it is to quit. Like the other players, whether she will take the best action or mistakenly quit depends on her own digit.

As we go backwards in the game and analyze the behavior of players b and c (in the eyes of the non-confused type of the previous player), we see that the probability of continuation, which is the probability of an odd number of mistakes, increases. Since the probability of an odd number of successes in n independent draws tends to $1/2$ as the number of draws grows to infinity, we know that at some point the probability will be above $\frac{1}{d} < \frac{1}{2}$. With our $\varepsilon = 0.1$, this first happens when we reach player c, for whom the probability is $0.243 > 0.2$. Now, all the types of player d assign a probability greater than 0.2 to the event that player c continues: for some the probability is 0.243, and for the others it is $1 - 0.243$. This means that the best action for d is to continue, rather than to quit, and all the types of d agree on that. Whether they will indeed continue or mistakenly quit depends on their own digit.

Now, the conclusion that the best action for d is to continue is common to all the types of all the previous players (e, f, ...). Therefore, to analyze the behavior of players d, e, f and so on, we can look at a simplified game. In this game, the last player is d, and her payoff from continuing is higher than that of quitting. The payoffs of the other players are unchanged.

The analysis of the simplified game goes as follows. Player d's best action is to continue, and thus she continues unless confused in her own shoes. Also players e and f

continue unless they are confused an odd number of times. Thus, the probability that a player continues are $1-0.1$ for d, $1-0.82$ for e, and $1-0.243$ for f.

What about player g? All her types assign probability either $1-0.243$ or $1-(1-0.243)$ to the event that f continues. Since both numbers are above 0.2, they all agree that the best action for g is to continue. Thus, whether g continues depends only on whether she is confused in her own shoes – probability $\varepsilon = 0.1$. Thus, we can again simplify the game, make g the last player, with the payoff to continuing higher than that of quitting (as we did with player d).

We therefore see that the probability of continuation follows cycles of length $n(d, \varepsilon) = 3$. The first cycle is different from the others, as the best action for the last player a is to quit, while for the simplified games corresponding to the next cycles the best end-action is to continue. Thus, the probabilities of continuation in the first cycle are 0.1, 0.18 and 0.243; for the other cycles we have the complementary probabilities 0.9, 0.82 and 0.757.

The following graph depicts the probability of continuation in the eyes of a type $t_{n+1} = 111\ldots1$ as a function of the number of generations from the end, for the parameters of the example in the Section 2 ($d = 5,\ \ \varepsilon = 0.1$).

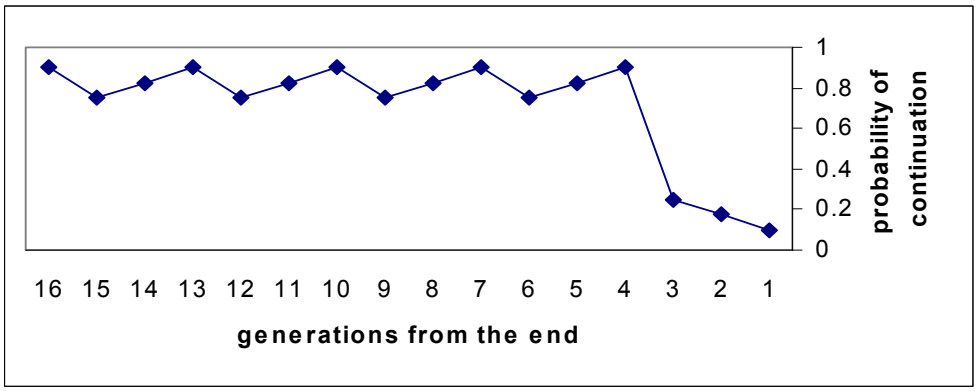

Figure 3: Continuation probabilites in the Centipede Game

4 Concluding Remarks

We introduced a small deviation from full common knowledge of rationality: players ascribe some small probability to the possibility that other players reach a different conclusion when they examine the same decision problem. We analyzed the centipede game and saw that the usual backward-induction, non-cooperative solution might not be robust to such a perturbation of the model.

4.1 Interpretation of the type space

In our type space there is common certainty that each of the players (1) believes that other players are sometimes confused (relative to her own point of view), and (2) is sure that she is right in all her calculations. 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 $111\ldots1$ who 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 $111\ldots1$. However, even the confused types mistakenly believe that they are always correct.

Both interpretations above are not compatible with a common-prior, Bayesian-game framework. Can our model be adapted to this more standard assumption? How would our analysis change if we assume that each type also doubts her own judgement?

The critical difference is that if a player doubts her own reasoning, she should better try to gain information regarding her best action from past decisions of other players, as these decisions were influenced by their analysis of her current decision problem. Differently put, since she understands that her own thinking only gives her a noisy signal regarding the correct computation, she should consider the preceding players' actions as additional signals that improve the accuracy of her own one. But now the analysis becomes very different. Since the analysis of a subgame depends on the history that

precedes it, forward-induction arguments come into play (even though the game is in agent-form): If a player knows that her decision to continue will affect the (sequential equilibrium) beliefs of the next agent, she must take this effect into account. In particular, if by continuing she can convince the next player to continue, then she should continue even if she believes that the best action for the next player is to exit.[4]

While studying the forward-induction consequences of a Bayesian framework could be interesting, we preferred to adhere, in this paper, to backward-induction reasoning (by which cooperation in the centipede game is perhaps more surprising), and leave this issue for future research.

4.2 Two-player centipede games and forward-induction

In our agent-form analysis, types were ascribed to agents. In a two-player alternating-moves centipede game, in contrast, a type is ascribed to a player. It is natural to assume that the type of the player should not change as the game proceeds. In other words, the way a player predicts the continuation of the game from some future node onwards, is the same at any decision node of hers that precedes that future node.

One may conjecture that this assumption implies a positive correlation between the *actions* taken by a given type at different node. That is, if a type continues at some node, it is likely that it also continues at its previous decision node. (This conjecture is based on the structure of the centipede game, in which a belief that the probability of continuation in the future is higher implies a higher incentive to continue in the present.)

The condition, that there is positive correlation between the actions of a type at different nodes, is the driving force behind forward-induction arguments. For example, in Kreps [1990], there is a "crazy" type who always continues. Thus, when a player is called to play, she knows that the opponent continued at the previous node, and this increases her assessment of the probability that the opponent will continue at the next node. This type of reasoning leads to cooperation at the beginning of long centipede games. Under the above conjecture, that our belief structure leads to a positive correlation

[4]We thank Asher Wolinsky for this observation.

in actions, one could view our framework as providing a possible foundation for the positive correlation assumption. That is, the positive correlation of actions becomes a result rather than an assumption.

4.3 Other games: the prisoners' dilemma

One may wonder about other canonical games. An important example is the finitely repeated prisoners' dilemma. One "lesson" from the forward-induction literature (such as Kreps et al. [1982]) was that there is a similarity between the centipede game and the finitely repeated prisoners' dilemma, as in both the same slight perturbation leads to an equilibrium with cooperation for a long while. Are the two games similar also under our perturbation? Or maybe the non-cooperation, backward-induction outcome in the finitely repeated prisoners' dilemma is robust to the introduction of small mistakes in reasoning?

It turns out that the analysis of that game is quite complicated, as the number of possible histories expands exponentially with the number of iterations. However, we can analyze a simplified version which is perhaps closer to the way most people think about the repeated prisoners' dilemma.

Consider a finitely repeated prisoners' dilemma game, in which each player has a different agent who plays at each stage. This game differs from the canonical agent-form since the same agent plays at all the nodes corresponding to a specific stage; nonetheless, with full common knowledge of rationality, backward induction yields the usual noncooperative solution. Now assume that each agent ascribes a small probability ε to the event that another agent *always* reaches the opposite conclusion when she analyses the same decision problem. That is, an agent suspects that any other agent might be "mistakenly wired" and always reach the "wrong" conclusion (from the first agent's perspective). Moreover, assume that these doubts are common knowledge.

This simplified model is very easy to analyze. Clearly, at the last stage both players defect (if they don't make a mistake). In the one before-the-last stage, a player who carried out a correct analysis will certainly defect. Moreover, a player who mistakenly expects cooperation in the last stage will also choose to defect. Although she errs in her

analysis of the last stage, she does not expect her action to affect her opponent's action in the next stage. Thus, she chooses the action that maximizes her stage payoff. The same reasoning applies to all the stages of the game. Players who do not err in their analysis of the current stage choose to defect, whatever doubts they may have regarding the opponent's reasoning. Thus the probability of cooperating is simply the probability ε that a player errs in her own shoes, when analyzing the current stage. In other words, in contrast to the centipede game, in the prisoners' dilemma mistakes do not accumulate: although the player may be mistaken in many places when she analyses the continuation of the game, the probability she will take the wrong action remains small. Therefore, if the mutually ascribed probability of mistakes ε is small enough, then no matter how many times the game is repeated, a player who is never mistaken will always defect.

Thus, if we think of the model of slight mistakes in reasoning as a *refinement* of subgame perfect equilibrium, the finitely repeated prisoners' dilemma differs from the centipede game. The subgame perfect no-cooperation equilibrium of the finitely repeated prisoners' dilemma is robust to the introduction of mutual doubts regarding the faultlessness of the players. This is not the case with the (long) centipede game in which a slight deviation from common knowledge of rationality yields a qualitatively different backward-induction outcome.

One possible objection to the above modeling of mistakes in the prisoners' dilemma, is that we only allow players to mistakenly choose fixed actions, rather than also allow for mistakenly choosing a history-dependent strategy. Clearly, a player will choose to cooperate only if she believes that the opponent who plays after her has a history-dependent strategy, such as tit-for-tat; if she believes her opponent's action is fixed – whether it is cooperation or defection – she will not cooperate. Maybe we could have restored cooperation in the prisoners' dilemma had we allowed mistakes in beliefs about history-dependent strategies?

It is easy to see that the answer is negative. Assume for example that a type, at the one-before-last stage, mistakenly believes that tit-for-tat is the best action at the last stage. If not confused in her own shoes, she will decide to cooperate – not to play tit-for-tat. Going back to an agent at the two-before-last stage, if she mistakenly believes that

the best action at the *last* stage is tit-for-tat, but given this mistake correctly computes the action at the one-before-last stage – which is "cooperate", then her best action is to defect! The only case in which she would not play defect is if she made a mistake at the one-before-last stage (or in her own shoes).

In other words, in the prisoners' dilemma game, the possibility of mistakenly assigning to players history dependent strategies with 1-recall (like tit-for-tat), can only affect the optimal action for an agent one stage earlier, but not further back in the game. That is, the only cases in which a player at stage k from the end might not defect are (i) if she made a mistake in her own shoes, or (ii) if she made a mistake in the shoes of the players at the next stage $k-1$. Further than that, a single mistake does not affect her best action. As a result, mistakes with probability of order ε do not accumulate to more than 2ε (ε times the bound on the recall $+1$).

The conclusion is that if we allow for mistakes to history-dependent strategies with recall bounded to 1, then, as the probability of mistake approaches 0, play in the repeated prisoner's dilemma converges to all-defect, regardless of the length of the game. The same line of reasoning applies also if we allow mistakes to history-dependent strategies with longer recall, as long as the bound on the recall is commonly known. Thus, even when we allow for history-dependent mistakes with bounded recall, we may still say that, unlike in the centipede game, the backward induction outcome in the finitely repeated prisoners' dilemma is robust to the inclusion of mistakes in reasoning of the kind we analyzed.

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:

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

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

The following two lemmata will be useful for the sequel.

Lemma A.1. *All the types in $E_{m+1}^{\ell,m}$ assign the same probability to $E_m^{\ell,m}$, all the types in $O_{m+1}^{\ell,m}$ assign the same probability to $O_m^{\ell,m}$, and these two probabilities are the same.*

Proof. For every type $t_{m+1} \in T_{m+1}$, denote by $P_{t_{m+1}} : T_m \to T_m$ the permutation of T_m that for each $t_m \in T_m$ inverts (i.e., changes 0's to 1's and 1's to 0's) all the entries of t_m corresponding to entries where t_{m+1} is confused (has entry 0). (That is, for $t_m = (t_m^n)_{n=1}^m$, each t_m^n is unchanged if $t_{m+1}^n = 1$ and inverted if $t_{m+1}^n = 0$.)

Note that, by definition, for every $A \subseteq T_m$

$$b_{t_{m+1}}^m(A) = b_{\mathbf{1}_{m+1}}^m\left(P_{t_{m+1}}(A)\right).$$

Note further that for any $t_{m+1} \in E_{m+1}^{\ell,m}$ (i.e., the number of 0 entries between ℓ and m is even), $P_{t_{m+1}}$ maps both $E_m^{\ell,m}$ and $O_m^{\ell,m}$ to themselves. Similarly, for $t'_{m+1} \in O_{m+1}^{\ell,m}$, $P_{t_{m+1}}$ maps $E_m^{\ell,m}$ onto $O_m^{\ell,m}$ and vice versa. Denote by $\mathbf{1}_{m+1}$ the type $111\ldots1$ of player $m+1$. Then,

$$b_{t_{m+1}}^m(E_m^{\ell,m}) = b_{\mathbf{1}_{m+1}}^m\left(P_{t_{m+1}}(E_m^{\ell,m})\right) = b_{\mathbf{1}_{m+1}}^m(E_m^{\ell,m}) \quad , \quad b_{t_{m+1}}^m(O_m^{\ell,m}) = b_{\mathbf{1}_{m+1}}^m\left(P_{t_{m+1}}(O_m^{\ell,m})\right) = b_{\mathbf{1}_{m+1}}^m(O_m^{\ell,m})$$
$$b_{t'_{m+1}}^m(O_m^{\ell,m}) = b_{\mathbf{1}_{m+1}}^m\left(P_{t'_{m+1}}(O_m^{\ell,m})\right) = b_{\mathbf{1}_{m+1}}^m(E_m^{\ell,m}) \quad , \quad b_{t'_{m+1}}^m(E_m^{\ell,m}) = b_{\mathbf{1}_{m+1}}^m\left(P_{t'_{m+1}}(E_m^{\ell,m})\right) = b_{\mathbf{1}_{m+1}}^m(O_m^{\ell,m})$$

which imply that all $b_{t_{m+1}}^m(E_m^{\ell,m})$'s and $b_{t'_{m+1}}^m(O_m^{\ell,m})$'s are the same, and all $b_{t_{m+1}}^m(O_m^{\ell,m})$'s and $b_{t'_{m+1}}^m(E_m^{\ell,m})$'s are the same. ■

Lemma A.2. *Suppose that all the types of m in $E_m^{\ell,m}$ take the same action, and all the types in $O_m^{\ell,m}$ take the opposite action. Let*

$$p = b_{\mathbf{1}_{m+1}}^m(E_m^{\ell,m})$$

1) *If* $p < \frac{1}{d}$ *or* $p > 1 - \frac{1}{d}$*, then all the types of* $m+1$ *in* $E^{\ell,m+1}_{m+1}$ *take the same action and all the types of* $m+1$ *in* $O^{\ell,m+1}_{m+1}$ *take the opposite action.*

2) *If* $\frac{1}{d} < p < 1 - \frac{1}{d}$*, then all the types of* $m+1$ *with* 1 *in entry* $m+1$ *continue, and all the types of* $m+1$ *with* 0 *in entry* $m+1$ *quit. In particular, similarly to case 1, all the types of* $m+1$ *in* $E^{m+1,m+1}_{m+1}$ *take the same action and all the types of* $m+1$ *in* $O^{m+1,m+1}_{m+1}$ *take the opposite action.*

Proof. In case 2, all the types of $m+1$ assign probability of more than $\frac{1}{d}$ that m continues. This is because if the types in $E^{\ell,m}_m$ continue, then by lemma A.1 all the types in $E^{\ell,m}_{m+1}$ assign to $E^{\ell,m}_m$ the same probability as the one assigned by $\mathbf{1}_{m+1}$, which is $p > \frac{1}{d}$, and the other types of $m+1$, those in $O^{\ell,m}_{m+1}$, assign to $E^{\ell,m}_m$ probability $1-p > \frac{1}{d}$. Similarly, if the types in $O^{\ell,m}_m$ continue, all the types in $E^{\ell,m}_{m+1}$ assign to $E^{\ell,m}_m$ probability $1-p > \frac{1}{d}$, and the types in $O^{\ell,m}_{m+1}$ assign to $E^{\ell,m}_m$ probability $p > \frac{1}{d}$. Thus, all the types of $m+1$ who are not confused in node $m+1$ continue, and all those who are confused there quit.

In case 1, denote by $C^{\ell,m}_m$ the set of types of m who continue – either $E^{\ell,m}_m$ or $O^{\ell,m}_m$. Now, $b^m_{\mathbf{1}_{m+1}}(C^{\ell,m}_m)$ is either p or $1-p$, and thus is either (a) less than $\frac{1}{d}$ or (b) more than $1-\frac{1}{d}$. by lemma A.1, in case (a) all the types of $m+1$ in $E^{\ell,m}_{m+1}$ assign to $C^{\ell,m}_m$ a probability smaller than $\frac{1}{d}$ and the types in $O^{\ell,m}_{m+1}$ assign to $C^{\ell,m}_m$ a probability larger than $1-\frac{1}{d}$; in case (b) all the types of $m+1$ in $O^{\ell,m}_{m+1}$ assign to $C^{\ell,m}_m$ a probability smaller than $\frac{1}{d}$ and the types in $E^{\ell,m}_{m+1}$ assign to $C^{\ell,m}_m$ a probability larger than $1-\frac{1}{d}$.

Denote by $Q^{\ell,m}_{m+1}$ the set of types of $m+1$ who assign to $C^{\ell,m}_m$ probability smaller than $\frac{1}{d}$, and by $C^{\ell,m}_{m+1}$ the set of types of $m+1$ who assign to $C^{\ell,m}_m$ probability larger than $1-\frac{1}{d}$. By the argument above, either $Q^{\ell,m}_{m+1} = E^{\ell,m}_{m+1}$ and $C^{\ell,m}_{m+1} = O^{\ell,m}_{m+1}$ (case a), or $Q^{\ell,m}_{m+1} = O^{\ell,m}_{m+1}$ and $C^{\ell,m}_{m+1} = E^{\ell,m}_{m+1}$ (case b).

To maximize expected utility, the types of $C^{\ell,m}_{m+1}$ who are not confused in node $m+1$ continue there and the types of $C^{\ell,m}_{m+1}$ who are confused in node $m+1$ quit. In contrast, the types of $Q^{\ell,m}_{m+1}$ who are not confused in node $m+1$ quit there, and the types of $C^{\ell,m}_{m+1}$ who are confused in node $m+1$ continue. So altogether, the types of $m+1$ who continue constitute either the set $E^{\ell,m+1}_{m+1}$ or the set $O^{\ell,m+1}_{m+1}$ and the types who quit constitute the

other. ■

Proof of Theorem 3.1. In node 1 (the last), type 1 (which constitutes the set $E_1^{1,1}$) quits and type 0 (which constitutes the set $O_1^{1,1}$) continues. With $\ell = m = 1$ we are thus in case 1 of lemma A.2, because $p = 1 - \varepsilon > 1 - \frac{1}{d}$. The conclusion of the lemma in this case is that the premise of the lemma obtains also with $\ell = 1$ and $m = 2$. Inductively, we can apply iteratively Lemma A.2 case 1, with $\ell = 1$ and increasing m, until p first falls below $1 - \frac{1}{d}$ (this must happen since the probability p of getting confused an even number of times decreases and tends to $\frac{1}{2}$). Let n be the minimal m for which $p < 1 - \frac{1}{d}$. Evidently, the smaller ε and d are, the larger is n. Then, we can apply case 2 of Lemma A.2. (since $\frac{1}{2} < p < 1 - \frac{1}{d}$) With $\ell = 1$ and $m = n$.

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

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.

Finally, to compute the lower bound on the probability of cooperation for players $i > n$ (as perceived by type 111...1 of player $i + 1$), note that at each iteration, if this probability is less than $1 - \frac{1}{d}$ we go to case 2 of lemma A.2., and the probability jumps back up to $1 - \varepsilon$. The lower bound is thus obtained if we take the lowest bound on probability for which case 1 of lemma A.2. applies, which is $1 - \frac{1}{d}$, and compute the probability after applying case 1 again. Now,

$$\begin{aligned} b^{m+1}_{\mathbf{1}_{m+2}}(E^{\ell,m+1}_{m+1}) &= (1-\varepsilon)\cdot b^{m}_{\mathbf{1}_{m+1}}(E^{\ell,m}_{m}) + \varepsilon \cdot b^{m}_{\mathbf{1}_{m+1}}(O^{\ell,m}_{m}) \\ &= (1-\varepsilon)\cdot b^{m}_{\mathbf{1}_{m+1}}(E^{\ell,m}_{m}) + \varepsilon \cdot \left(1 - b^{m}_{\mathbf{1}_{m+1}}(E^{\ell,m}_{m})\right) \end{aligned}$$

Why? With probability $(1 - \varepsilon)$ the new digit $(m + 1)$ is 1, and in this case every type in $E^{\ell,m}_m$ becomes a type in $E^{\ell,m+1}_{m+1}$, and with probability ε the new digit is 0, and in this case every type in $O^{\ell,m}_m$ becomes a type in $E^{\ell,m+1}_{m+1}$. Substituting $1 - \frac{1}{d}$ for the lowest

bound of $b^m_{\mathbf{1}_{m+1}}(E^{\ell,m}_m)$, we obtain:

$$b^{m+1}_{\mathbf{1}_{m+2}}(E^{\ell,m+1}_{m+1}) \geq (1-\varepsilon)\left(1-\frac{1}{d}\right)+\varepsilon\frac{1}{d} = 1-\frac{1}{d}-\varepsilon\left(1-\frac{2}{d}\right)$$

■

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] Ben-Porath, E. 1997. "Rationality, Nash Equilibrium and Backward Induction in Perfect-Information Games," *Review of Economic Studies* 64:23-46.

[3] McKelvey, R.D. and T.R. Palfrey 1992. "An Experimental Study of the Centipede Game," *Econometrica* 60:803-836.

[4] Kreps, D. 1990. *A Course in Microeconomic Theory,* Princeton University Press.

[5] 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.

[6] Rosenthal, R.W. 1981. "Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox," *Journal of Economic Theory* 25:92-100.

[7] Selten, R. 1982. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Form Games," *International journal of Game Theory* 4:25-55.