Backwards induction in the centipede game

John Broome & Wlodek Rabinowicz

The game

Imagine the following game, which is commonly called a 'centipede game'. There is a pile of pound coins on the table. X and Y take it in turns to take either one or two coins from the pile, and they keep the coins they take. However, as soon as either of them takes two coins, the game stops, and the rest of the coins are cleared away. So long as they each take only one coin when their turn comes, the game continues till the pile is exhausted. Suppose the number of coins is even, and X has the first turn.

Assume both X and Y aim to maximise their own gain only, and they are rational throughout the game. We mean 'rational' to imply only that they believe the logical consequences of their beliefs, and that they do not choose an option if there is some other available option that they believe would give them more money. For the moment, assume they have a common belief in rationality throughout the game. That is to say, throughout the game, they each believe that each of them will be rational throughout the game, that each of them will believe throughout the game that each of them will be rational throughout the game, and so on.

Traditional argument

Given these assumptions, a standard backwards-induction argument concludes that X will take two coins at her first move, thereby ending the game. The argument is as follows.

Suppose the game gets to the point where there are only two coins on the table. It is X's turn, and she will take both coins. That way she gets both of them, whereas she believes that if she only takes one, Y will take the second. Taking both is the only rational thing for X to do, so she does it because she is rational.

Next suppose the game gets to the point where there are only three coins on the table. It is Y's turn. Y believes that, if he takes only one of the coins, at the next round X will take the remaining two. We have established that taking two will be the only rational thing for X to do, and by assumption Y believes X will only do what is rational. So Y will take two himself and thereby end the game. That way, he gets two of the three coins on the table, whereas otherwise he gets only one. Taking two is the only rational thing for him to do, so he does it because he is rational.

Analysis 59.4, October 1999, pp. 237–242.

Now suppose the game gets to the point where there are only four coins on the table. It is X's turn. X believes that, if she takes only one of the coins, at the next round Y will take two and thereby end the game. We have established that taking two will be the only rational thing for Y to do, and by assumption X believes Y will only do what is rational. So X will take two herself and thereby end the game. That way, she gets two of the four coins on the table, whereas otherwise she gets only one. Taking two is the only rational thing for her to do, so she does it because she is rational.

And so on. We can conclude that X takes two coins at the first round.

This reasoning has been impugned (by, amongst others, Binmore (1987), Reny (1987), Pettit and Sugden (1989); for a contrary view, see Sobel 1993). It depends on assuming the players maintain a common belief in rationality throughout the game. But that is a dubious assumption. Suppose X was to take only one coin in the first round; what would Y think then? Since the backwards-induction argument says X should take two coins, and it is supposed to be a sound argument, rationality apparently requires her to take two. So when she takes only one, Y might be entitled to doubt her rationality. Alternatively, Y might doubt that X believes Y is rational, or that X believes Y believes X is rational, or Y might have some higher-order doubt. X's deviant first move might cause a breakdown in common belief in rationality, therefore. Once that goes, the entire argument fails.

The argument also assumes that the players act rationally at each stage of the game, even if this stage could not be reached by rational play. But it is also dubious to assume that past irrationality never exert a corrupting influence on present play.

Assumptions

However, the backwards-induction argument can be reconstructed for the centipede game on a more secure basis.[1] It may be implausible to assume a common belief in rationality throughout the game, however the game might go, but the argument requires less than this. The standard idealisations in game theory certainly allow us to assume a common belief in rationality at the beginning of the game. It also allows us to assume this common belief persists so long as no one makes an irrational move. That is enough.

[1] This article contains a simplified version of one of the arguments that are set out in more detailed and precise form, using rather weaker assumptions, in Rabinowicz 1998. Versions of the reconstructed argument appear in Hern 1998 and in Binmore 1996. Aumann1998, which also contains a version of the argument, is mentioned at the end of this article. An early sketch of the argument appears in Sugden 1991..

More precisely, we assume:

(0) At each round in the game that has been reached without any irrational move, the player at that round acts rationally.
(1) At each round in the game that has been reached without any irrational move, the player at that round believes (0).
(2) At each round in the game that has been reached without any irrational move, the player at that round believes (1).

And so on.

Rysiek Sliwinski has shown us an objection to assumptions (1), (2) and so on. Suppose one player makes a move that is actually rational, but that the other player believes is irrational. Then the game will arrive at the next round without any irrational move's having been made, but even so, the second player may no longer believe in the rationality of the first. To counter this objection, we can derive (1), (2) and so on from assumptions that seem definitely consistent with the traditional idealisations in game theory. They are:

(A) At the beginning of the game, both players have no false beliefs.
(B) During the game, both players acquire only beliefs that are true.
(C) Both players retain all their beliefs so long as they are consistent with their acquired beliefs.
(D) At the beginning of the game, there is a common belief in (0), (A), (B) and (C).

From (A), (B) and (C), it follows that both players retain throughout the game all the beliefs they have at the beginning, because true beliefs are consistent with true beliefs. Since by (D) they believe (0) at the beginning of the game, they believe it throughout the game. In particular, (1) follows. (1) follows from (A), (B), (C) and (D), all of which both players believe at the beginning of the game, by (D). (A common belief implies a belief in itself.) Since both players believe the consequences of their beliefs, they believe (1) at the beginning of the game. Since they retain all their beliefs, they believe (1) throughout the game. In particular, (2) follows. And so on.

So we think (1), (2) and so on are acceptable assumptions. We also assume that the player at any round has correct beliefs about what moves have previously been made, and furthermore about what move he or she makes at that round itself. This second clause implies that the rationality of a move is determined by the beliefs the player has at the moment of choosing, rather than beforehand.[2]

[2] Rabinowicz' second proof in his (1998) offers an alternative to this 'at-choice' perspective on rationality.

Proof

Now the argument. Notice first that, if any particular round in the game is reached, (0) implies it is reached without any irrational move. By (0), X acts rationally at the first round, so if a second round is reached, it is reached without any irrational move. Therefore, if a second round is reached, Y acts rationally there, by (0), so if a third round is reached, it is reached without any irrational move. And so on.

Notice second that, if any particular round in the game is reached, the player at that round believes it has been reached without any irrational move. If any round is reached, we have shown it is reached without any irrational move. Therefore the player at that round has a 'level (1) belief', as we shall call it: he or she believes (0). It follows that he or she believes the consequence of (0) demonstrated in the previous paragraph, including the consequence that this particular round has been reached without any irrational move.

Now suppose the game gets to the point where there are only two coins on the table. This can only have happened without any irrational move's having been made. So by (0), X acts rationally. Given that, for the same reason as before, she takes both coins.

Next suppose the game gets to the point where there are three coins on the table (the 'three-coin round'). This can only have happened without any irrational move's having been made. So Y acts rationally at this round, and also has a level (1) belief. On this basis we shall prove Y takes two coins.

As a hypothesis for reductio, suppose Y takes only one coin at the three-coin round. By assumption, Y believes he makes this move at this round. Since he has a level (1) belief, he believes both that all the previous moves have been rational, and that he acts rationally at this round. So he believes the game will arrive at the next round (the 'two-coin round') without any irrational move's having be made. Y therefore believes X will act rationally at the two-coin round; this is implied by Y's level (1) belief. We have just seen that, if X acts rationally at the two-coin round, it follows that she will take both remaining coins. Y's level (1) belief implies X will do this. So, given the hypothesis that he takes only one coin at the three-coin round, Y believes at that round that this one coin is all he will get. On the other hand, he also believes that, if he were to take two coins instead, he would get two. His taking one coin is therefore not rational, contrary to (0). So the hypothesis must be false: if the game gets this far, Y takes two coins.

Next suppose the game gets to the four-coin round. This can only have happened without any irrational move's having been made. So X acts rationally at this round, and also has both a level (1) and a level (2) belief: she believes that, at any move that is reached without an irrational move,

each player acts rationally and has a level (1) belief. On this basis we shall prove X takes two coins.

As a hypothesis for reductio, suppose X takes only one coin at the four-coin round. By assumption, X believes she makes this move at this round. Since she has a level (1) belief, she believes both that all the previous moves have been rational, and that she acts rationally at this round. So she believes the game will arrive at the three-coin round without any irrational move's having been made. X therefore believes Y will act rationally at the three-coin round and have a level (1) belief at this round; this is implied by X's level (1) and level (2) beliefs, respectively. We have just seen that, if Y acts rationally at the three-coin round and has a level (1) belief at this round, it follows that he will take two coins then. X's level (1) and level (2) beliefs imply Y will do this. So, given the hypothesis that she takes only one coin at the four-coin round, X believes at that round that this one coin is all she will get. On the other hand, she also believes that, if she were to take two coins instead, she would get two. Her taking one coin is therefore not rational, contrary to (0). So the hypothesis must be false: if the game gets this far, X takes two coins.

And so on. We conclude that X will take two coins in the first round, and finish the game.

Comments

Having reached this conclusion, it is tempting to ask: what would happen if X took only one coin in the first round? This line of questioning has led to some interesting discussion (for example in Binmore 1996 and in Sugden 1991). However it cannot falsify our conclusion that the backwards induction solution follows from our assumptions, provided our argument is valid. If you want to object, you must either object to the assumptions or to the logic of the argument.

A warning: We have argued for backwards induction in the centipede game, and the argument can be immediately extended to all games where the move that is recommended by backwards induction at any round terminates the game at that round (cf. Rabinowicz 1998). But we have not been able to extend it further than that.

Robert Aumann has proved a similar conclusion for the centipede game. His assumptions differ from ours in several respects; for one thing he assumes common knowledge of rationality rather than common belief. But he similarly avoids the dubious assumptions that apparently underlie the standard version of the argument for backwards induction. Expressed roughly and converted to our own terms, Aumann's remarkable proof is this. As a hypothesis for reductio, suppose there is a solution to the centi-

pede game that does not terminate at the first round. Consider this solution, or if there is more than one such solution, consider the one that continues the longest. This solution will end with one of the players – let it be Y, but it does not matter which – taking two coins. Now consider the previous round, where X takes just one coin. From the perspective of this round, the game will end at the next round. This fact follows from the players' common belief in rationality. Since X shares the common belief, and believes the consequences of her beliefs, she believes the game will end at the next round. But then it is irrational for her to take only one coin. This contradicts that the game will continue to the next round. Therefore the hypothesis is false. So the game will end at the first round.

This proof is in effect an elegant abridgement of ours. We believe ours is more transparent because it spells out how X acquires the belief that the game will end at the next round.[3]

The University, St Andrews
Fife, KY16 9AL, UK
john.broome@st-andrews.ac.uk

Lund University
Kungshuset, 222 22 Lund, Sweden
wlodek.rabinowicz@fil.lu.se

References

Aumann, R. 1998. A note on the centipede game. *Games and Economic Behavior* 23: 97–105.

Binmore, K. 1996. Rationality and backward induction. Typescript.

Binmore, K. 1987. Modelling rational players: part 1. *Economics and Philosophy* 3: 179–213.

Hern, R. 1997, *Rational Choice Theory When Tastes are Changing Through Time*. PhD thesis. University of Bristol.

Pettit, P., and R. Sugden. 1989. The backward induction paradox. *Journal of Philosophy* 86: 169–82.

Reny, P. 1988. *Rationality, Common Knowledge and the Theory of Games*. PhD thesis. Princeton University.

Rabinowicz, W. 1998.Grappling with the centipede. *Economics and Philosophy* 14: 95–126.

Sobel, H. 1993. Backward induction arguments in finitely iterated prisoners' dilemmas: a paradox regained'. *Philosophy of Science* 60: 114–33.

Sugden, R. 1991. A rational choice: a survey of contributions from economics and philosophy. *The Economic Journal* 101: 751–85.

[3] Our thanks to Howard Sobel and Rysiek Sliwinski for helpful comments. This note was written while John Broome was a Visiting Fellow at the Swedish Collegium for Advanced Study in the Social Sciences. He thanks the Collegium for its hospitality.